DOCUMENT RESUME

ED 201 663 TM 810 256

AUTHOR Drahozal, Edward C.
TITLE In Defense of Norms for School Averages and Developmental Standard Score Scales.
PUB DATE 30 Jan 81
NOTE 15p.: Paper presented at the Southwest Educational Research Association (January 30, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Educational Practices; Elementary Secondary Education; *Local Norms; *National Norms; *Scores; *Standardized Tests; *Test Interpretation; *Test Norms
IDENTIFIERS *Variability

ABSTRACT

This paper is divided into two parts. The purpose of Part I is the review of some of the history of school average norms, beginning with Dr. E. F. Lindquist's call for such norms in 1948; the development of such norms for standardized achievement tests; current policies and practices regarding aggregated school data; and a presentation of norms data for pupils and school averages for two tests with markedly different within-grade to between-grade variability. Part II briefly discusses various procedures used to obtain developmental standard score scales; provides information about developmental standard score scales used by five publishers; responds to the comment, "who needs developmental standard score scales, anyway?"; and discusses the validity of the most common criticisms of developmental standard score scales. The point is made that no one score so epitomizes perfection that it is "the" score and "the" answer to all problems associated with test interpretation and use. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED201663

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

In Defense of Norms for School Averages

and

Developmental Standard Score Scales

Edward C. Drahozal
Riverside Publishing Company

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. C. Drahozal
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Southwest Educational Research Association
January 30, 1981

TM 800 257

Norms for school averages are available for some tests and test batteries and should be used by school personnel. However, very few school districts with low-scoring schools report percentile ranks for school averages whereas most school districts with a preponderance of above average schools report percentile ranks for school averages. Why? Quite simply, although school personnel are forced to provide parents with scores for pupils which are at the first percentile, full disclosure rules and regulations do not require disclosure of the fact that, compared with other schools in the country, a local school is at the first percentile. Conversely, personnel in a school district with above-average schools are more than happy to report that, compared to all schools in the country, one or more of their schools is at the 99th percentile.

This double-standard for reporting exists despite the fact the Standards for Educational and Psychological Tests includes a standard that reads:

> D7. Where it is expected that a test will be used to assess groups rather than individuals (i.e., for schools or programs), normative data based on group summary statistics should be provided. Essential
>
> Comment: For example, it is inappropriate to evaluate schools by using norms developed for the evaluation of individuals. It is also inappropriate to compute group means for nonlinear scales such as percentile ranks derived for individual norms. p.24

In Frederick G. Brown's recent NCME publication, Guidelines for Test Use: A Commentary on the Standards for Educational and Psychological Tests, page 79, additional comments are made supporting the use of school average norms.

Thus, in reporting the average results of their several hundred schools, a large city test director reports the unbelievable information that no school in the district is below the eighth percentile and not one school in the district is above the ninetieth percentile. Compared with what?

The purposes of the first part of this paper are (1) to briefly review some of the history of the school average norms, beginning with Dr. E.F. Lindquist's call for such norms in 1948; (2) the development of such norms for standardized achievement tests; (3) current policies and practices regarding aggregated school data; and (4) a presentation of norms data for pupils and school averages for two tests with markedly different within-grade to between-grade variability.

I will read a few excerpts from a speech Dr. Lindquist made in 1948. Lin was rather soft-spoken, on most occasions, and perhaps the tenor of the first paragraph was such that some people in the audience did not believe that norms for school averages are that important. The entire speech is in my paper. The speech was first published, I believe, in 1966 in

the American Council on Education book entitled, Testing Problems in Perspective, pages 269-271, which was edited by Anne Anastasi.

"What I have to say is not at all profound, and I should be able to say it in just two or three mintues. As elementary as the observation I have to make, however, judging from current test standardization practices and from the literature on testing, it is one that is sorely in need of stressing even in as sophisticated a group as this one.

What I wish to do is to put in a plug for a special type of norm for educational achievement tests. It is a type of norm which, in consideration of the major uses generally made of achievement tests, should always be provided with tests of that character, but one which in actuality is very rarely provided.

Let me remind you first of two major uses which we generally claim for educational achievement tests. First, we claim that these tests will provide the teacher and counselor with a more adequate and dependable basis for the educational guidance of the individual pupil and for the individualization of instruction. Second, we claim that these tests will be useful to the school administrator as a partial basis for evaluating the content and methods of instruction, or in curriculum evaluation in general.

Now with reference to the first of these purposes, quite obviously we need norms that are descriptive of the distributions of measures of achievement of pupils. We interpret a pupil's score by comparing it with the scores of other pupils in the relatively homogeneous population to which he belongs. The second purpose, however, is the evaluation of group, not of individual, achievement. What the principal wants to know is, "How are we doing as a school?" or "How does this grade group compare in average achievement in relation to other similar grade groups?" Quite obviously, what is needed here is a type of norm that is descriptive of distributions of measures of group achievement, i.e., of distributions of school averages, not of distributions of individual pupil scores.

Since such norms are rarely provided, what the school principal does is to attempt to interpret his grade averages or grade medians in terms of the norms that are provided--norms descriptive only of distributions of pupil scores. What he really does, without often knowing that he is doing so, is to assume that the school averages are distributed just as are individual pupil scores. School averages, of course, are much less variable than pupil scores (although not as much less as many may assume), and he immediately makes wrong interpretations of his results on that account. A school average that coincides with the 75th percentile in the pupil score distribution, for example, might be a remarkably high average, but since it is "at the 75th percentile," it might not be regarded as high by the principal. What he does would not be so serious, however, if there were a constant relationship between the variability of school averages and the variability of pupil scores. In that case he could at least tell on what tests his school did well in relation to other schools, and could rank the tests in order of the relative performance of his school, even though he would be misled as to the true relative performance on any one test. Actually, however, the relationship is not constant, but varies widely from subject to subject and from one grade level to another.

I have with me here some data concerning the relative variability of individual pupil scores and of school means for 34 widely used achievement tests. I am not going to read all of these data to you but will just give you a sample of them. They are all based on large numbers of cases--in no instance upon less than 193 schools, in some cases as many as 400, and in no case less than 6,000 pupils--so the results are quite stable so far as sampling errors are concerned.

The tests include tests of elementary school achievement, such as the Iowa Tests of Basic Skills, and tests of high school achievement, such as the Cooperative subject achievement tests.

In the first place, let me say that, in every case, the variability of school means is far larger than would be the variability of means of random samples of the same size drawn from the same total population. If, in any case, one were to make a simple analysis of the total variance into "between schools" and "within schools" components, one would find that the "between schools" variance is from two to ten times as large as that which could be accounted for by random sampling alone. The ratio would vary greatly, however, from test to test.

In third-grade arithmetic (Iowa Tests of Basic Skills) the ratio of the 10th to the 90th percentile range for school averages and for pupil scores is exactly .5. For sixth-grade arithmetic it jumps to .55, but for third-grade language it is .36. For a high school physics test the ratio is .61, while for a United States history test at the same grade level it is .37. In general, the ratio seems to be largest where the opportunity to learn what is tested is relatively restricted to the classroom and to the formal school curriculum and to be least where the achievement measured might more often result from out-of-school experiences and informal learning.

Well, I think I have said enough to make my point. School averages are less variable than pupil scores, but much more variable than the means of random samples. More important, the relative variability of school averages differs markedly from test to test and grade to grade. If school administrators, then, are to interpret properly measures of <u>school</u> achievement for purposes of curriculum evaluation or for the evaluation of group achievement in general, they must by supplied with norms specifically adapted to that purpose--i.e., they must have norms descriptive of distributions of school <u>averages</u> for defined populations of schools, just as for the interpretation of individual achievement they now have norms descriptive of distributions of pupil scores for defined populations of pupils."

How are such norms developed? Quite simply, by obtaining frequency distributions of school average data from the national standardization program, plotting and smoothing these distributions, and reading the percentile ranks which correspond to various averages. I suspect that some school average norms are obtained from user data after a standardization using a pre-prepared model which simulates the national "school population." I have no quarrel with such a procedure provided, of course, that the model is developed before, not after, the school data are in-house.

What are current policies and procedures? Some schools use school average norms as a matter of long-standing policy. Some administrators, such as the large city administrator mentioned in my introduction, use pupil norms because that is the type of norm reported by several publisher's. A few people report neither pupil nor school average norms. Several report both!

Some people report percentile ranks of median pupil scores. I suppose that as long as one does not attempt to deceive someone into believing that such a score provides good information about relative achievement levels among schools, this practice will continue to grow. Strictly speaking, in order for such data to be meaningful, a table of norms of school medians would be necessary. However, no such tables exist, to my knowledge. In fact, I believe only Riverside Publishing Company and Addison-Wesley-ETS provide norms for school averages for their most recent test batteries. One publisher, SRA, provided such norms for previous editions of the tests and it appears as if they provide such norms on sample reports I have seen for their 1978 SRA Achievement Series; however, I cannot locate any specific information in the SRA materials or literature available to me to confirm this.

Among the people who do or do not report school average norms are a group that report scores at several percentile levels or percentages of pupils by score ranges, nationally and locally. I do not think this ought to be a substitute for school average norms. Such information complements school average norms and tells school board members why a school is at the 1st or 99th percentile compared to other schools and why X-amount of money is needed for gifted-talented programs and/or programs for educationally handicapped students.

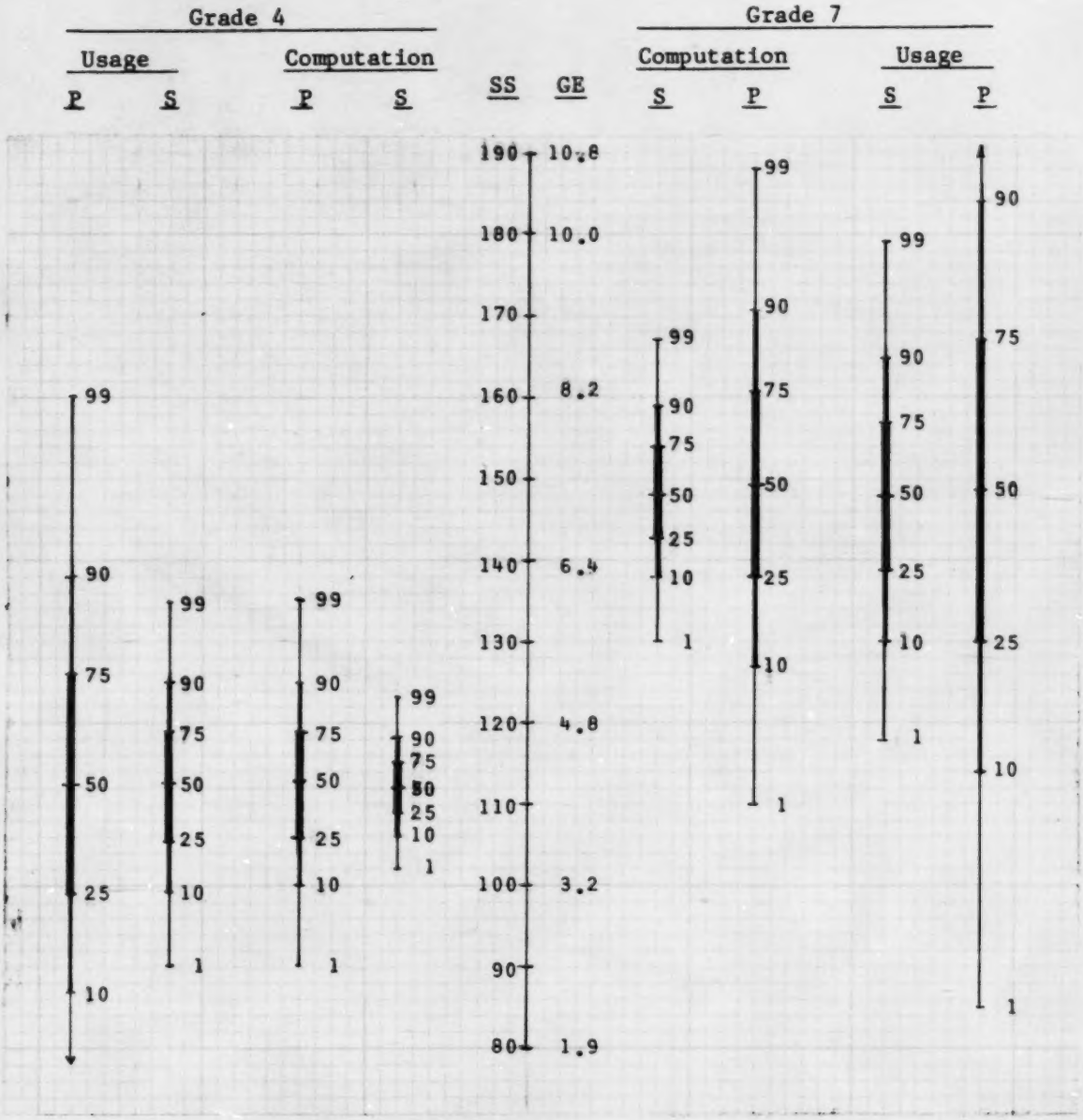
Lindquist mentioned that school average norms vary by grade and test. The figure below shows the pupil and school average score ranges for Grades 4 and 7, Fall Norms for the Iowa Tests of Basic Skills Language Usage test and the Math Computation test. The usage test is a good example of a test that is affected both positively and negatively by outside influences; math computation is learned almost exclusively in school and is typically taught in "lock-step" fashion.

FIGURE 1 is on Page 5

Note that in Grade 4, pupils (P) are about half as variable in Math Computation as they are in Language Usage. Also note that the variability of schools (S) on usage is almost exactly the same as the variability of pupils on computation. In Grade 7 the relationships between school average norms and pupil norms are similar to Grade 4, but pupil variability in computation exceeds school variability in usage. Last, the language usage ratios are greater than the language ratios to which Lindquist referred in 1948, but they are still less than the math ratios.

FIGURE 1

ITBS Language Usage and Math Computation, Grades 4 and 7,
Pupil (P) and School Average (S) Scores for
Selected Percentile Ranks



The purposes of the second part of the paper are (1) to briefly discuss various procedures used to obtain developmental standard score scales, (2) to provide information about developmental standard score scales used by five publishers; (3) to respond to the comment, "Who needs developmental standard score scales, anyway?" made by a well-known latent-trait practitioner at a recent AERA-NCME convention; and (4) to discuss the validity of the most common criticisms of developmental standard score scales.

The first two purposes of part two are directly related to each other. Different publishers use different vertical scaling procedures. The "numbers" assigned to developmental score scales differ from publisher to publisher. The notion that "a standard score is a standard score is a standard score" that seems to pervade the profession needs to be dispelled.

How are developmental standard score scales developed?

There are two data sources--the frequency distributions of the weighted national standardization sample and the data from the vertical scaling study. For some tests the vertical equating data come from the national standardization. However, for most of the current achievement batteries, the relationships among levels are obtained from separate vertical scaling studies and the score scale is established by using some variation of the venerable Thurston Absolute Scaling technique. Most scales are established by assigning a mean and a standard deviation, somewhat arbitrarily it seems, at a given grade level and then "chain-linking" upward and downward to establish a K-12 scale. A few scales are established by assigning means at two grade levels and then jointly normalizing all within-grade distributions for each test on a common K-12 score scale.

Although the national standardization samples differ greatly in size and somewhat in terms of stratification procedures, all of the major publishers appear to use three basic stratifications: region of the country; school population; and some index of socioeconomic level of the school district such as income-education, Title I enrollments, and/or free/reduced rate school lunch participation. The vertical scaling is most often done by randomly assigning students to two or three groups and administering different levels of the test to each group. A better procedure, but one that is less often used, is to develop scaling tests which transcend several levels and administer these scaling tests to students at several levels with the regular battery.

What are the characteristics of various publishers' developmental standard score scales?

I'll begin with the Iowa Tests of Basic Skills-Tests of Achievement and Proficiency (ITBS/TAP) score scale. The ITBS/TAP scale was developed by using the 2-point method. The Grade 3 Fall mean for all subtests was assigned a value of 100 and the Grade 8 Fall mean was assigned a value of 160. The range of scores from K-12 is approximately 15 to 250.

Now I'll go alphabetically! The California Achievement Tests (CTB-1977) scale score (sometimes called an expanded scale score) was anchored in the spring using Grade 10 (Level 19), with a mean of 600 and the standard deviation of 90, and then linked downward to Level 10, Kindergarten. The scores range from about 10 to 950.

Metro '78 (Psych Corp - 1978) standard score scale ranges from 1 at the Preprimer Level to 999+ at the Advanced II Level. The score scale for each test was established in the fall by anchoring at the level with the greatest variability--usually the Elementary Level. The mean and SD used vary somewhat from test to test.

The SRA Achievement Test (SRA - 1978) Growth Scale Value (GSV) was developed in 1968 using data from the 1963-64 standardization. All tests were anchored in Grade 4, Fall, with a mean of 250 and a standard deviation of 50. The scale was linked up and down from Grade 4. The current scale is linked to the original scale via a series of equating studies and is not a pure lily-white, equal-interval scale that some mathematical statisticians desire. So what! Scores range from about 20 to 780.

The CIRCUS-Sequential Tests of Educational Progress III (ETS/Addison-Wesley - 1974-1979) use a "standard score scale" which was anchored in Grade 12, Fall, using Level J, with a mean of 300 and a standard deviation of 20. The score scale ranges from about 200 to 370 and is available from Spring Grade 1 (CIRCUS, Level C) through Grade 12 (STEP III, Level J).

Table 1 provides some statistics for these five batteries for the Reading Comprehension and Mathematics Concepts subtests. (For Metro '78 the data are for Reading and Mathematics; CAT '77 has a math concepts/applications subtest. Also, STEP has Grade Level Indicators (GLI's) not GE's!) Are the score scales similar? The answer is that they are not. Granted, the point at which scaling begins and the score scale used affect the relationships among means and standard deviations. Yet, even if one takes these factors into consideration, there are some very real differences. For example, for one battery the math variability is equal to or greater than the variability in reading. For two batteries anchored similarly, various ratios among standard deviations that one would expect to be similar are quite dissimilar.

TABLE 1

Reading Comprehension and Mathematics Concepts
Means and Standard Deviations
Grades 4 and 7, Fall

	Grade 4						Grade 7			
	*GE Units (Years)		*Standard Score Units				Standard Score Units			
	Reading		Reading		Concepts		Reading		Concepts	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
ITBS/TAP	4.25	1.42	113.0	17.9	113.8	15.7	148.9	23.1	149.1	19.8
CAT	4.5	1.89	431.0	53.1	376.0	41.6	521.0	71.2	508.0	67.2
METRO	4.1	2.4	666.4	73.8	576.8	90.2	735.2	82.9	712.4	82.0
STEP III	4.4	*2.1	265.6	15.1	260.8	13.0	280.0	16.9	277.8	16.0
SRA	4.62	2.27	283.0	46.4	268.3	39.5	336.7	64.3	329.1	56.9

*GE data are provided as reported in the manuals. Standard score data for some test batteries were rounded to the nearest tenth. For STEP the GLI data for Grade 4 were estimated. The same program was used to estimate the Standard Score data for Grade 4. The SS mean and SD for Grade 4 Reading were 265.4 and 15.2, respectively, which are nearly identical to the reported 256.6 and 15.1.

Since standard score scales are difficult to compare, GE data for Grade 4 Reading is provided. (It is recognized that GE's are synonyms with anathema in some quarters.) The GE data indicate that most of the Grade 4 GE distributions are skewed positively. But why are the GE SD's for the various batteries so different? The tests? Vertical scaling? The standardization sample? Because these are GE's? The latter is not the reason for most of the batteries. As one bit of evidence one could submit the standard score means and medians for these tests! The reader can speculate why. Or perhaps dig through the norms tables and other statistical data to find the answer.

The fact is that the various developmental score scales are quite different in appearance, within-grade variability, and the amount of grade-to-grade overlap. Which is better? You know which one I think is better. My caution to you is to look at the score scales relative to the uses which must be made or may have to be made of the test results. I'll cite two examples.

Suppose your school reading supervisor wants to use the reading score as a indicator of instructional reading level. Let's first assume that the GE was to be used as an indicator of instructional reading level. Table 2 provides data on reading for Fall, Grade 5, from three test batteries. The data were derived using equipercentile procedures, under the assumption that the standardization samples of all three tests were similar, an assumption that was more or less borne out by the Anchor Test Study. The GE's for each test are to the right of the selected PR's. The IRL GE on the left is the average of the instructional reading levels for two Metro '78 levels appropriate for use in Grade 5 and a Grade 5 ISR obtained from a study done by Dr. MacGinitie with the Second Edition of the Gates-MacGinitie Reading Test. The IRL's were obtained using somewhat different procedures and the 3.5 GE at PR of 5 seems to be a bit high. (However, the overall "reading range" for 90% of the students is slightly greater than four years, which is about what reading experts believe to be reasonable.) Also, the median is slightly above 5.1 or 5.2. The data speak for themselves. At the lower percentile ranks the differences among tests are not nearly as great as at the top!

TABLE 2

Instructional Reading Levels and Grade Equivalent Scores for Selected Percentile Ranks for the Reading Subtests of Three Test Batteries

IRL GE	PR	Test 1 GE	Test 2 GE	Test 3 GE
7.8	95	9.5	8.0	12.8
7.2	90	8.4	7.4	10.4
6.2	75	6.7	6.4	7.5
5.3	50	5.1	5.2	5.0
4.5	25	3.8	4.1	3.3
3.7	10	2.7	3.2	2.5
3.5	5	2.2	2.8	2.2

Of course, the scapegoat for these sorts of inconsistencies is the GE scale. Clearly it is not the function of the scale. All score scales used for a given test/test battery are derived from the same standardization program and the same vertical scaling study. The GE is "impure" (or amoral) according to the purists. The standard score is pure(r), for most tests. And, of course, since slightly different statistics are used to develop each score scale, some slight differences between the GE and the standard score are to be expected.

In order to relate the standard scores for each test, all of which have different characteristics, I first normalized the ISR using a mean of 100 and an SD of 20. The using the grade-to-grade overlaps of the ISR as a rough baseline, the standard scores for each of the test batteries were similarly normalized. (It is recognized that the Test 3 distribution is not "normal." However, it has exactly the same characteristics as the published data. E.g., the range from PR50 to PR95 is 19% greater than the range from PR50 to PR5, which is the same as the published norms.)

The data in Table 3 are similar to the GE data in Table 2.

TABLE 3

Normalized Standard Scores for Instructional Reading Levels and for Reading Tests of Three Test Batteries

IRL	PR	Test 1	Test 2	Test 3
133	95	142	135	150
126	90	134	128	143
114	75	118	115	124
100	50	100	100	100
86	25	82	85	80
74	10	67	71	64
67	5	59	64	58

For the second example, let's use a not so atypical scenario. The superintendent of schools calls you in and says, "The school board has told me that, beginning this fall, we will test all students in Grade 10 to ascertain which students are below a Grade 8.5 proficiency level in reading and math. Then we will remediate these kids and, except for the real eight-balls, we will get as many as possible above that Grade 8.5 level by the end of Grade 12!" You go back to your office and, since your district is very, very average relative to the nation, you use standard scores equal to the 50th percentile in Grade 8 to estimate the percent of pupils in your district who are likely to score below Grade 8.5. You go back to the superintendent with the figures in the left hand column of Table 4. The superintendent says, "Wow, I don't have much money because when the school board told me we had to do this, I didn't think that we would have to remediate anywhere near 25% of the students. But, since Test 1 has the fewest students, let's take Test 1."

The light bulb turns on and you say, "I think you ought to take Test 2. We will have a few more students in the program, but we ought to be able to show that the remedial program really works if we use Test 2. Let me show you what I mean." To which the superintendent replies, "All you'll do is show me a bunch of statistics. I don't give a darn about these and neither does the school board. They want results with the least amount of money. So let's take Test 2."

TABLE 4

Percent of Students Who Score Below the Grade 8 Mid-Year Median, Reading and Mathematics

	Beginning of Grade 10		End of Grade 12	
	Reading	Math	Reading	Math
Test 1	25%	25%	18%	19%
Test 2	28%	25%	24%	24%
Test 3	38%	35%	22%	22%

Were you, the test director, correct? Of course, simply by retesting the students selected in Grade 10 you would expect fewer than 24% to score below a score equal to a Grade 8.5 level! Of course, if there is someone on the school board who is really bright, both the test director and the superintendent would lose their job.

These two examples show quite conclusively that there are differences in developmental standard score scales. The differences are probably due, for the most part, to differences in the vertical scaling procedures. Since such differences exist, a third question, "Why even bother with developmental standard score scales?" is often raised.

The answer is quite simple. Teachers, students, parents, the school board, and administrators are interested in growth as well as status. Thus, some sort of score scale which shows development is demanded. Profiles such as the ITBS Pupil Profile or the SRA Growth Scale Profile are used to show growth to students, parents, and others.

Ten or twelve years ago some people thought they would beat the development issue by using CRT's and reporting percent improvement for items, objectives, domains, or what have you. What happened? Within one year parents, school boards, and others asked, "How good is that?" or "Compared to what?" So we now have national norms and developmental score scales for CRT's; and some of these score scales are really quite poor.

Last, what are some of the common criticisms of developmental standard scores?

1. Developmental standard score scales are difficult or impossible to interpret. Thus, they are "inadequate as a measure of relative standing" and are not developmentally descriptive.

This is not exactly true. Psychometrists and statisticians who constantly study standard score scales sometimes do a commendable job of interpreting standard scores. Teachers, administrators, and lay persons seldom understand

much about standard score scales but this may be a blessing in disguise. First, these non-technicians obviously cannot misunderstand such a scale. Some understanding is a prerequisite for misunderstanding and misinterpretation. Second, because the scale is not understood, non-technicians are forced to use scales such as percentile rank, percentile bands, stanines, or similar scores which ought to be used. If a parent asks about growth, a teacher could refer to the standard score as a "mathematically derived scale" and then show the parent a growth scale chart that indicates Sheila or Sean grew from 150 to 160 in Reading, and that 10 points growth is above, below, or about average growth for the time period in question. For example, if the student was "average" and obtained standard scores of 150 and 160 in Grades 4 and 5, respectively, and ten points is "normal" or average growth for between these grades at this score level, one could say to a parent that the student's growth was about average.

2. That's another problem with developmental standard scores. The growth differs by test, grade, and score level for every test I know. Thus, you can't compare standard scores for different subtests across grade level and at different score levels. They mean different things in different subjects at different grades.

Research in learning has repeatedly shown that growth (educational development) differs by subject matter, grade level, ability level, and other factors. Thus, one would expect that any score scale linked to grade-to-grade growth will differ by subtest, grade level, and score level. Low-scoring or low-ability students tend to grow less than high-scoring or high-ability students. As Lindquist stated in 1948, grade-to-grade development and, therefore, grade-to-grade overlap differ for science, math, social studies, reading, language arts, and study-skills. To give one example, growth in reading is very rapid in the lower grades and growth in science is not, whereas the opposite is true, especially for above average students in Grades 7-12.

Some developmental standard score scales are developed in such a manner that average or near-average scores are similar for all tests in a given grade for Grades 2 through 9. However, even for such score scales, extreme scores---above the 80th percentile or below the 20th percentile---differ across tests and cannot be directly compared on any developmental score scale, whether it's a standard score scale or a grade equivalent score scale. All grade-based scales have different standard deviations across tests and grades.

3. What you are saying is that developmental standard scores are grade-related. That's another problem with them.

There are two common denominators available to persons who develop score scales, age and grade. Age equivalents or standard scores by age group are often used in non-graded situations or for infants and very young people. They are also used with ability and aptitude tests which measure characteristics not directly related to an instructional program. However, since age requirements for entering school differ from state-to-state and by region of the country and since learning is most directly related to the amount of instruction provided the learner, grade is the best common denominator for any developmental score scale. Grade equivalents have always been by definition, related to median performance in a grade or a succession of grades. Standard scores are developed by using grade means.

4. The highest and lowest developmental standard score values are really mathematical estimates of very low or very high performance and are rarely based on actual testing of students at that level of performance.

No achievement test or test battery developed for use in schools is administered to a representative sample of, for example, two year old children or twenty-five year old persons. Even if someone would do so, these people are not in the same sort of instructional mainstream as students in Grades K-12 or K-16; so scores derived at the extremes on these people probably would be worse than the mathematical estimates of extreme scores obtained by testing students in Grades K-12. The same criticism can be made of most grade equivalent scales, but it is really not terribly valid.

5. Because developmental standard score scales are jointly normalized across grades at a given time of the year, most are not equal-interval scales for the all grades and times-of-year. In addition, they characterize as normally distributed a lot of pupil distributions that are not at all normally distributed, such as reading skills in Grade 1 or math in Grade 12.

Both of these criticisms are accurate to a degree. However, a well-constructed standard score scale is much more "equal-interval," whatever that means to most people, than a well-constructed grade equivalent score scale, at least below Grade 3 and above Grade 8. Thus, one can provide averages, standard deviations, and other aggregated data that are somewhat mathematically correct than if similar data were calculated using grade-equivalent scores. In response to the second comment, you can't have your cake and eat it. If you want a better mathematical scale you have to live with a scale that is jointly normalized over the entire grade range, K-12, whether or not the distributions are normal. You can't stop at Grade 2 or Grade 10 just because the distributions below and above these grades don't meet a test of normality for all subjects or skill areas measured.

6. Developmental standard scores are used in schools to set standards of performance.

People, not scores, set standards.

7. Developmental standard scores are affected by differences in curriculum emphases from school system to school system.

So are raw scores, from which all scores are derived.

8. If developmental standard scores are so good, it seems that publishers ought to get together and give us a single basic score scale like the SAT, which has a mean of 500 and an SD of 100.

Just as there are differences in philosophy that are reflected in test content, test format, and even subtests, so are there differences of opinion in regard to score scales and factors which relate to score scales, such as vertical scaling procedures. Furthermore, even if a comoon score scale was agreed upon next year, ten years from now the average achievement level in the anchor grade(s) might be different for different subtests, the grade-to-grade overlap might be different, and other factors related to the score scale may change. Thus, the score scale agreed upon now would have different characteristics ten years from now, just as the Verbal, Mathematics, and Total scores of the SAT no longer have a mean of 500 an SD of 100.

To summarize, I hope at least some of you will use norms for school averages as part of your summary of aggregated test data, if for no other reason but to avoid being criticized by truth-in-testing advocates. Second, I hope the second part of this paper has provided sufficient information to help you understand that score scales are different and that no one score so epitomizes perfection that it is the score and the answer to all problems associated with test interpretation and use.